Measuring Inequality by Asset Indices: The case of South Africa

Martin Wittenberg and Murray Leibbrandt

UNU-WIDER conference

5 September 2014

Core Intuition

- Main methods of generating asset indices (PCA, Factor Analysis, MCA) look for correlations between different "assets"
  - Latent variable interpretation: what is common to the assets must be "wealth"
- This breaks down when there are assets that are particular to sub-groups (rural areas) such as livestock
  - These assets are typically **negatively** correlated with the other assets
- Resulting index will violate the assumption that people with a lower score always have less "stuff" than people with a higher score

Summary

- The way in which asset indices are created (e.g. in the DHSs) does things which are not transparent to users
  - The indices show anomalous rankings
  - They tend to exaggerate urban-rural differences
- It is possible to construct indices in a way which sidesteps these issues
- In the process it is possible to give a cardinal interpretation to the indices, i.e. we can estimate inequality measures with them
- When applying these measures to South African data we find that "asset inequality" has decreased markedly between 1993 and 2008
  - This contrasts with the money-metric measures
  - If incomes rise across the board then asset holdings with a static schedule will show increases in attainment while inequality will stay constant
- However, creation of asset indices should proceed carefully -- examining whether the implied coefficients make sense

Outline of the talk

- Motivation
- "Standard" approach for creating asset indices
- Some desirable principles for creating asset indices
- Thinking about asset inequality:
  - With one binary variable
  - With two binary variables
  - Multidimensional inequality
- Applying the approach to DHS data
- Evolution of Asset Inequality in South Africa 1993-2008
- Conclusions

Motivation

- Asset indices have become very widely used in the development literature, particularly with the release of the DHS wealth indices
  - 13 900 "hits" for "DHS wealth index" on Google Scholar
  - 2 434 Google Scholar citations of the Filmer and Pritchett article
  - 591 Google Scholar citations of the Rutstein and Johnson (DHS wealth index) paper
- Use of these indices has been externally validated (e.g. against income)
- But in at least some cases they are internally inconsistent (as we will show)
- Asset indices have proved extremely useful in broadly separating "poor" from the "rich“
- Cannot use indices to measure inequality or changes in inequality -- yet in some cases assets is all we have

Purpose of the paper

- Raise questions about the semi-automated way in which asset indices are produced
- Argue for an alternative method of calculating such indices
- Show that this method avoids some pitfalls, plus it enables the calculation of inequality measures
- These measures produce interesting insights when applied to S.A. data
- BUT we don't want to substitute one mechanical way of creating indices for another

Literature: Principal Components

- The Filmer and Pritchett (2001) paper argued that the first principal component of a series of asset variables should be thought of as "wealth".
- This interpretation has underpinned its adoption by the DHS as the default approach for creating the “DHS wealth index”

Latent variable interpretation

- Write asset equations as

$$a_1 = v_{11}A_1 + v_{21}A_2 + \cdots + v_{k1}A_k$$
$$a_2 = v_{12}A_1 + v_{22}A_2 + \cdots + v_{k2}A_k$$
$$\dots$$
$$a_k = v_{1k}A_1 + v_{2k}A_2 + \cdots + v_{kk}A_k$$

with $A_1, A_2 \dots, A_k$ mutually orthogonal

- Then A_1 is the variable that explains most of what is "common" to the assets a_i

The mechanics

- Variables are standardized (de-meaned, divided by their standard deviations)
- The scoring coefficients are given by the first eigenvector of the correlation matrix

Consequences:

- Asset indices have mean zero (i.e. can’t use traditional inequality measures on them)
- The implicit “weights” on each of the assets are a combination of the score **and** the standardization
  - Generally not reported/interrogated

Validation

- Filmer and Scott
  - Compare rankings according to different asset indices against each other
  - Compare to per capita expenditure
- Asset indices highly correlated with each other
- Somewhat highly correlated with per capita expenditure
  - Correlation highest where per capita expenditure well predicted by community characteristics etc
  - Where private goods (in particular food) not such a big component of per capita expenditure

Criticisms

- Index is intrinsically discrete
  - Can limit its ability to discriminate at the top/bottom of the distribution
  - Performs better if at least some “continuous” variables (rooms) are used
- Correlation between groups of binary variables constructed from categorical ones
- Should infrastructure variables be included? Can have independent impacts on outcome of interest

Some desirable principles for creating asset indices

- Monotonicity

$$\text{if } (a_1, a_2, \dots, a_k) \geq (b_1, b_2, \dots, b_k)$$
$$\text{then } A(a_1, a_2, \dots, a_k) \geq A(b_1, b_2, \dots, b_k)$$

Note: this presumes we are talking about “goods” not “bads”

- Absolute zero (desirable, not essential)

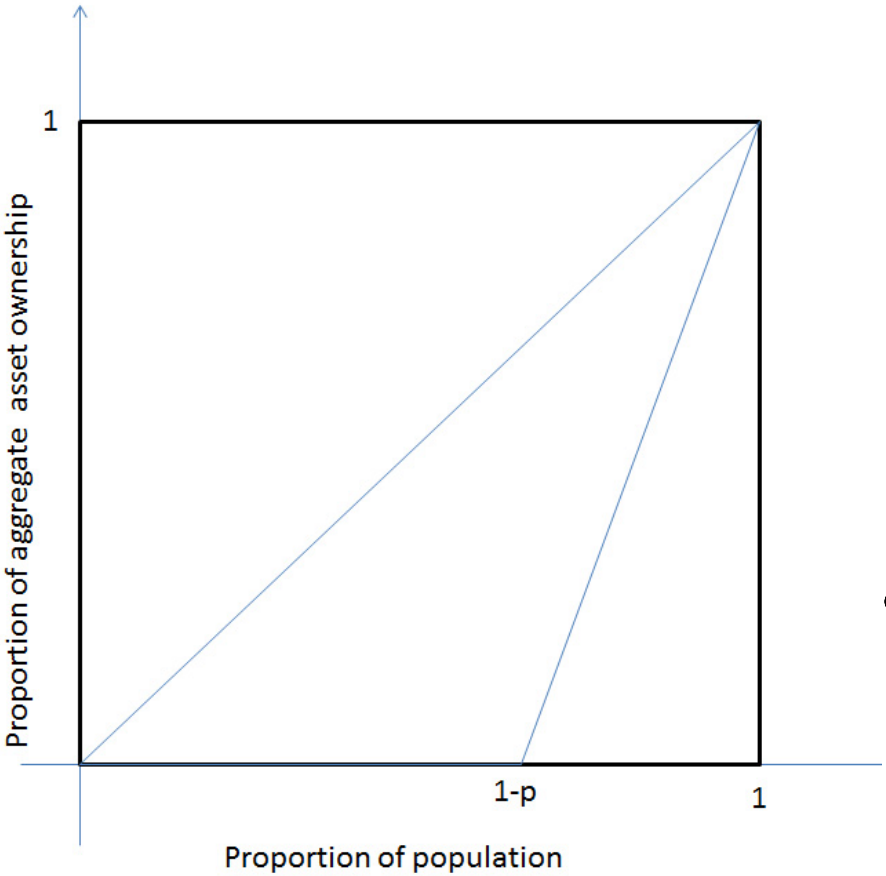
$$A(0,0, \dots, 0) = 0$$

- Robustness – should work whether or not the variables are continuous/binary

Thinking about inequality using binary variables

- Many of the traditional “thought experiments” don’t work in this context:
  - e.g. there is no way to do a transfer from a richer to a poorer person while keeping their ranks in the distribution unchanged
  - It is impossible to scale all holdings up by an arbitrary constant

The case of one dummy variable



- Plot the Lorenz curve
  - Gini coefficient is just $1-p$
  - Maximal inequality when p=ε
  - Decreases monotonically as p goes to one
- Similar view of inequality when using coefficient of variation

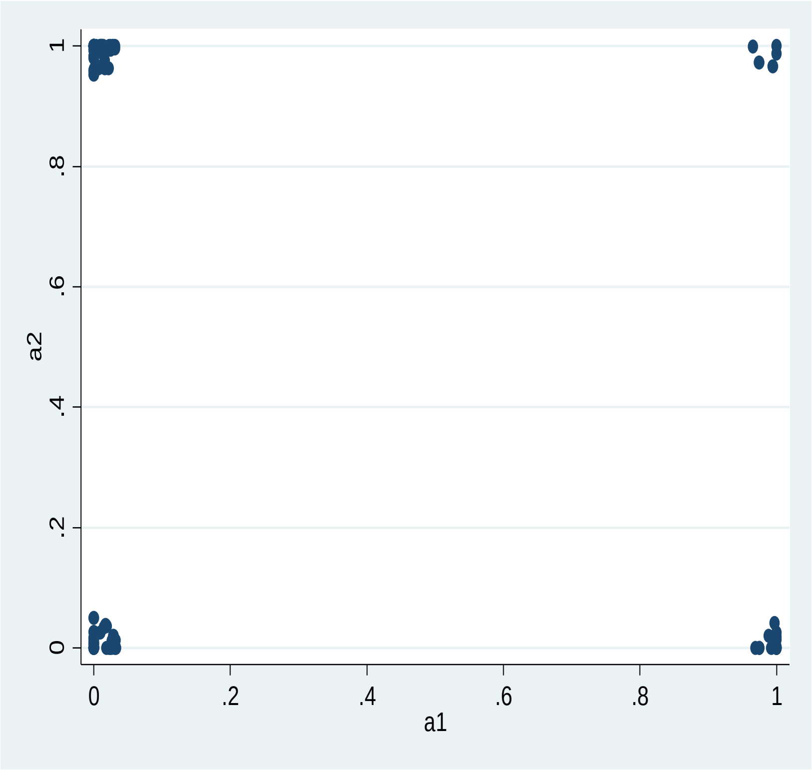
Two binary variables

- One additional complication that occurs when you have more than one variable is dealing with the case of a "correlation increasing transfer"
  - e.g. the asset holdings (1,0) and (0,1) versus (0,0) and (1,1)
- Most people would judge the second distribution to be more unequal than the first

PCA index

- We can derive expressions of the value of the PCA index as a function of
  - the proportions p_1 and p_2 who hold assets 1 and 2 respectively
  - and p_{12} the fraction who hold both
- The range (and the variance) of the index shows a U shape with minimum near p_1 (the more commonly held asset)
  - Unbounded near 0 and 1

More critically



- The assets will be negatively correlated whenever $p_{12} \leq p_1 p_2$
- In this case one of the assets will score a **negative** weight in the index

Why is this the case?

- The "latent variable" approach can make sense of the negative correlation only if one of the assets is reinterpreted as a "bad", e.g. a_1
- This will result in the rankings:
  $A(0,1) \geq A(1,1)$ and $A(0,0) \geq A(1,0)$
- Not hard to construct examples where (1,1) scores lower than (0,0)
- Is this relevant? – Yes! Empirical work

Multidimensional Inequality Indices

- Tsui: “Generalized entropy” measures
- Problem is that the theory assumes continuous positive (cardinal) variables

Banerjee's "Multidimensional Gini"

- Create an "uncentered" version of the principal components procedure:
  - Divide every variable by its mean (in the binary variable case p_i)
  - This makes the procedure "scale independent"
    - In the continuous variable case
  - It has the side-effect of paying more attention to **scarce assets** in the binary variable case
    - BUT this will also prove troublesome in some empirical cases
  - Then extract the first principal component of the cross-product matrix
- Calculate Gini coefficient on this index

What does this do?

- This procedure is guaranteed to give non-negative scores
- Banerjee proves that the Gini calculated in this way obeys (using continuous variables) obeys all the standard inequality axioms
- PLUS it will show an increase in inequality if a “correlation increasing transfer’’ is effected

In the case of asset indices

- It is guaranteed to give an asset index that obeys the principle of monotonicity
- It will have an absolute zero
- And it can be used to calculate Gini coefficients even when all variables are binary variables.

Application to the DHS wealth index

VARIABLES	DHS WI	UC PCA	UC PCA2	PCA	PCA2	MCA	FA
water in house	0.252***	0.209	0.565	0.708	0.707	0.329	0.289
electricity	0.180***	0.0814	0.220	0.663	0.657	0.300	0.265
radio	0.0978***	0.0515	0.140	0.467	0.477	0.206	0.113
television	0.160***	0.101	0.273	0.678	0.680	0.312	0.301
refrigerator	0.179***	0.136	0.369	0.735	0.738	0.343	0.413
bicycle	0.0923***	0.600	1.401	0.490	0.501	0.233	0.137
m.cycle	0.169***	52.57		0.788	0.821	0.412	0.193
car	0.175***	0.490	1.202	0.766	0.777	0.368	0.320
rooms	0.0102***	0.0176	0.0482	0.0977	0.105	CAT	0.0221
telephone	0.196***	0.378	0.989	0.813	0.818	0.387	0.397
PC	0.210***	4.984	14.42	0.967	0.982	0.481	0.296
washing machine	0.203***	0.654	1.696	0.870	0.877	0.421	0.452
donkey/horse	-0.0880***	2.836	4.523	-0.293		-0.118	-0.0849
sheep/cattle	-0.118***	0.291	0.509	-0.375		-0.156	-0.0909
Observations	11,666	12,136	12,136	12,136	12,136	12,136	12,136
R-squared	0.999	1.000	1.000	1.000	1.000	1.000	1.000

Comparing the PCA 2 and UC PCA2 rankings

Quantiles of UC PCA2	Quantiles of PCA 2					
	1	2	3	4	5	Total
1	2 368	482	0	0	0	2 850
2	530	1 145	748	0	0	2 423
3	34	429	1 277	586	0	2 326
4	0	66	275	1 463	399	2 203
5	175	104	55	84	1 912	2 330
Total	3 107	2 226	2 355	2 133	2 311	12 132

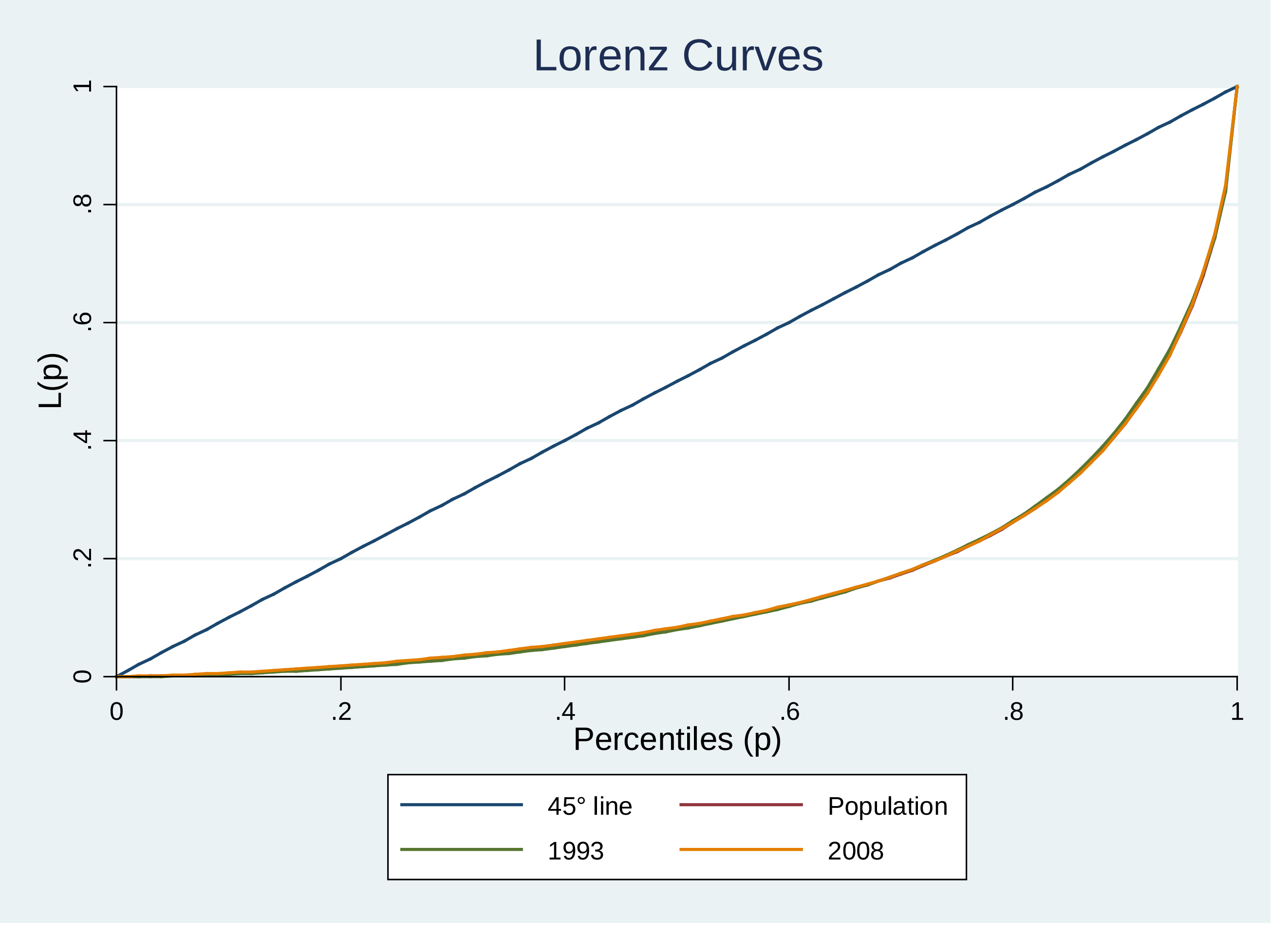
Proportion poor (bottom 40%)

		Linearized		
Over	Mean	Std. Err.	[95% Conf. Interval]	
DHS				
capital, large city	0.098	0.013	0.072	0.123
small city	0.178	0.024	0.131	0.225
town	0.204	0.031	0.142	0.265
countryside	0.720	0.020	0.681	0.759
PCA 2				
capital, large city	0.146	0.014	0.119	0.173
small city	0.220	0.021	0.179	0.261
town	0.291	0.032	0.229	0.353
countryside	0.648	0.019	0.610	0.686
UC PCA 2				
capital, large city	0.198	0.015	0.169	0.227
small city	0.275	0.022	0.232	0.317
town	0.372	0.033	0.308	0.437
countryside	0.597	0.016	0.566	0.628

Asset inequality by area

Group	Estimate	STE	LB	UB
1: capital, large city	0.566	0.009	0.549	0.583
2: small city	0.538	0.014	0.511	0.566
3: town	0.569	0.023	0.524	0.614
4: countryside	0.609	0.014	0.582	0.636
Population	0.623	0.007	0.610	0.636

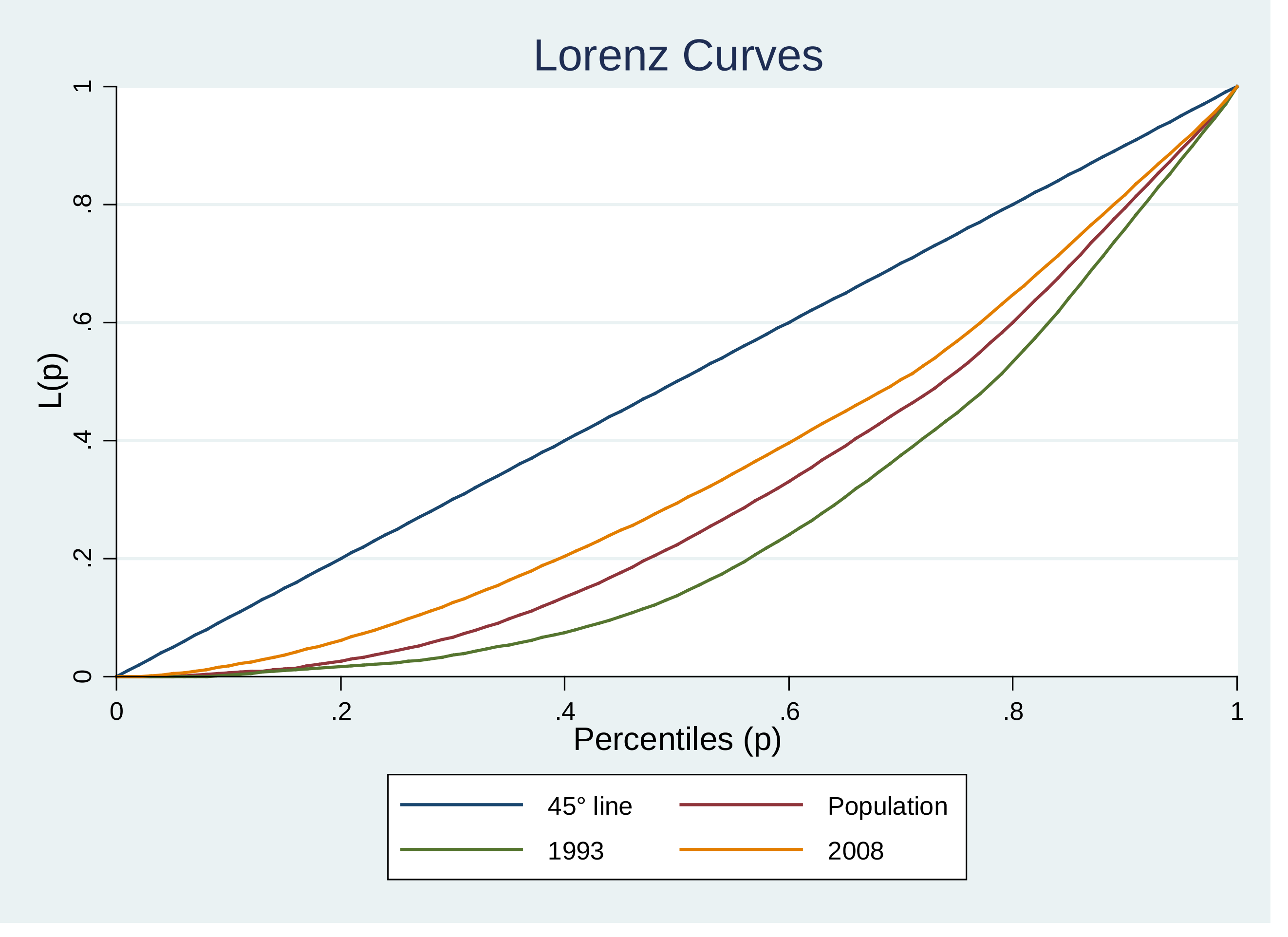
South Africa 1993-2008



Asset holdings

		Linearized		
Over	Mean	Std. Err.	[95% Conf. Interval]	
electricity				
1993	0.459	0.024	0.411	0.507
2008	0.779	0.020	0.740	0.818
pipedwater				
1993	0.506	0.027	0.454	0.559
2008	0.697	0.025	0.648	0.746
radio				
1993	0.811	0.008	0.796	0.826
2008	0.694	0.012	0.672	0.717
TV				
1993	0.477	0.018	0.441	0.512
2008	0.703	0.017	0.671	0.736
fridge				
1993	0.399	0.020	0.360	0.438
2008	0.609	0.020	0.569	0.648
motor				
1993	0.247	0.016	0.215	0.279
2008	0.220	0.018	0.184	0.256
livestock				
1993	0.110	0.011	0.089	0.132
2008	0.100	0.011	0.078	0.122
landline				
1993	0.242	0.018	0.206	0.278
2008	0.143	0.015	0.114	0.172
cellphone				
2008	0.807	0.011	0.786	0.828
phoneany				
1993	0.242	0.018	0.206	0.278
2008	0.827	0.010	0.808	0.847

South Africa - Assets



Why the difference?

- Incomes have increased across the board
  - Inequality stayed constant
- Asset register, however, is fixed:
  - Higher proportions of South Africans have access to these
  - Hence this measure goes down
- The two methods really ask different questions
  - Asset inequality measure looks at the gap between the "haves" and the "have nots"
    - Is scale dependent
  - Income inequality looks at the distribution of incomes, where essentially everyone has something
    - Is scale independent